

US005628003A

United States Patent [19]

Fujisawa et al.

[11] Patent Number: 5,628,003

[45] Date of Patent: May 6, 1997

[54] DOCUMENT STORAGE AND RETRIEVAL
SYSTEM FOR STORING AND RETRIEVING
DOCUMENT IMAGE AND FULL TEXT DATA

[75] Inventors: Hiromichi Fujisawa, Tokorozawa;
Atsushi Hatakeyama, Kokubunji;
Yasuaki Nakano, Hino; Junichi
Higashino, Koganei; Toshihiro
Hananoi, Naka-gun, all of Japan

[73] Assignee: Hitachi, Ltd., Tokyo, Japan

[21] Appl. No.: 111,511

[22] Filed: Aug. 24, 1993

Related U.S. Application Data

[62] Division of Ser. No. 139,781, Dec. 30, 1987, Pat. No.
5,265,242, which is a division of Ser. No. 894,855, Aug. 8,
1986, abandoned.

Foreign Application Priority Data

Aug. 23, 1985 [JP] Japan 60-184181

[51] Int. Cl.⁶ G06F 17/30

[52] U.S. Cl. 395/615; 364/225.4; 364/237.6;
364/282.1; 364/DIG. 1; 364/DIG. 2; 382/230;
382/306; 382/181

[58] Field of Search 395/600; 382/9,
382/57, 27

References Cited**U.S. PATENT DOCUMENTS**

4,404,649 9/1983 Nunley et al. 235/379
4,553,261 11/1985 Proessl 382/57
4,574,395 3/1986 Kato 382/61
4,607,385 8/1986 Maeda 382/9
4,616,854 10/1986 Landrum et al. 283/74
4,630,308 12/1986 Hongo 382/27
4,741,045 4/1988 Denning 382/9
4,748,678 5/1988 Takeda et al. 382/56
4,941,125 7/1990 Boyne 395/600

FOREIGN PATENT DOCUMENTS

58-144980 8/1983 Japan .
60-114967 6/1985 Japan .
60-105039 6/1985 Japan .

OTHER PUBLICATIONS

Information Processing Institute, "Data Base Machine", Jul.
15, 1980, Ohm Co.

Primary Examiner—Thomas G. Black

Assistant Examiner—Hosain T. Alam

Attorney, Agent, or Firm—Antonelli, Terry, Stout & Kraus,
LLP

[57] ABSTRACT

A document storage and retrieval system is provided with means for storing a document body in the form of image, means for storing text information in the form of a character code string for retrieval, means for executing a retrieval with reference to the text information, and means for displaying a document image relating thereto on a retrieval terminal according to the retrieval result. Such a form of the system is available for retrieving the full contents of a document and also for displaying the document body printed in a format easy to read straight in the form of image. Accordingly, users are capable of retrieving documents with arbitrary words and also capable of reading even such a document as is complicated to include mathematical expressions and charts through a terminal in the form of image, the same as on paper. Further, the invention provides a system wherein the text information for retrieval is extracted automatically from the document image through character recognition. Since a precision of the character recognition has not been satisfactory hitherto, a visual retrieval and correction have been carried out without fail by operators. However, there is no necessity for the operators to attend therefor according to the invention. Thus, the text information for retrieval can be generated at the cost of practical time and money even in case of volumes of documents.

11 Claims, 16 Drawing Sheets

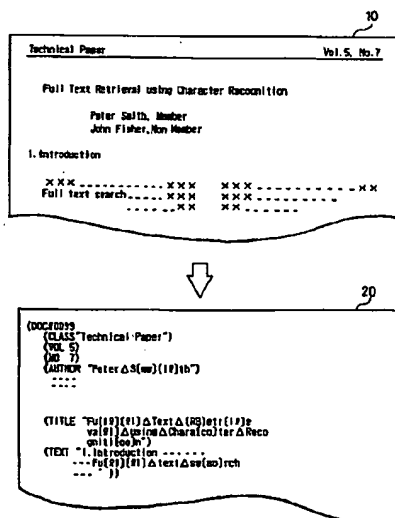


FIG. 1

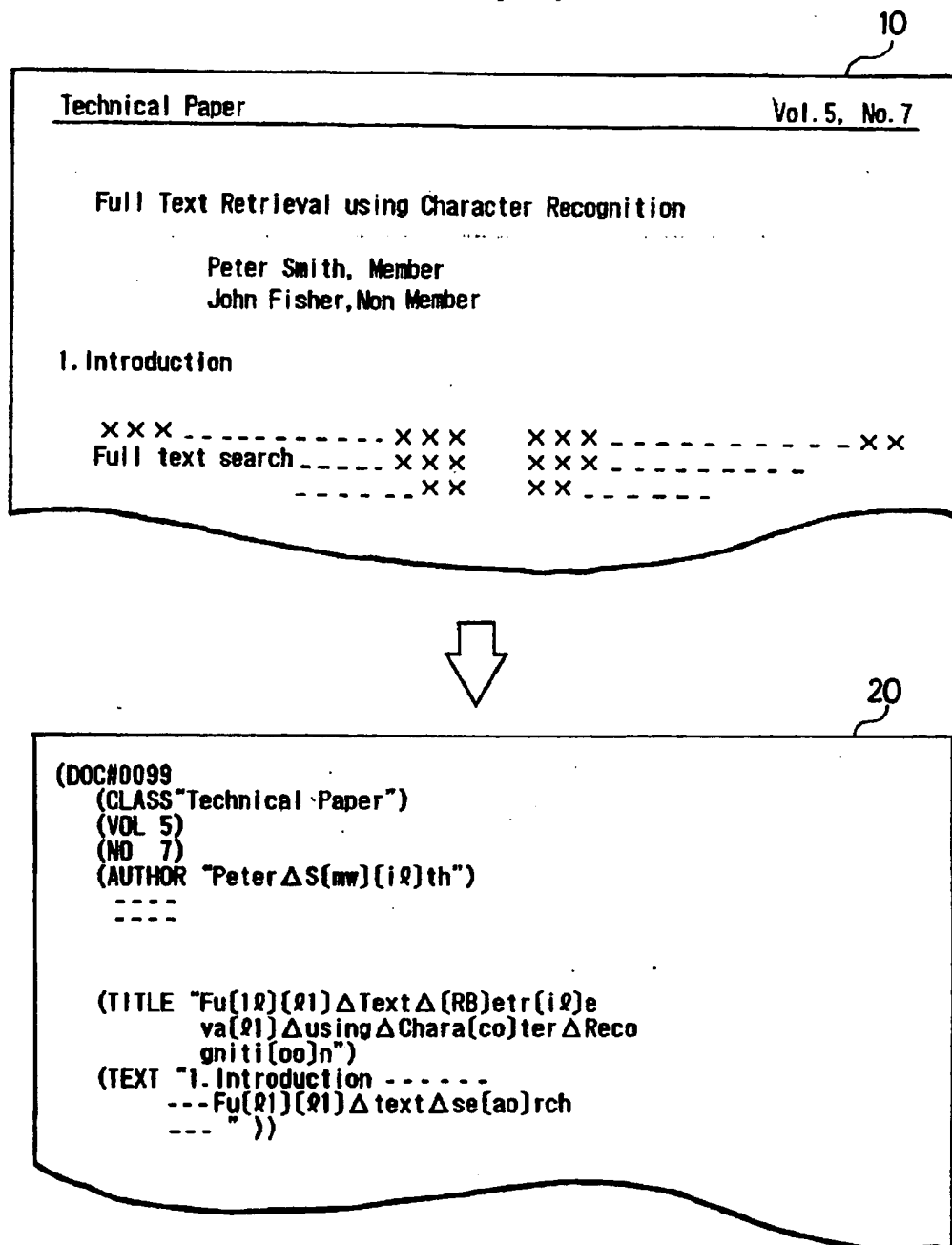


FIG. 2

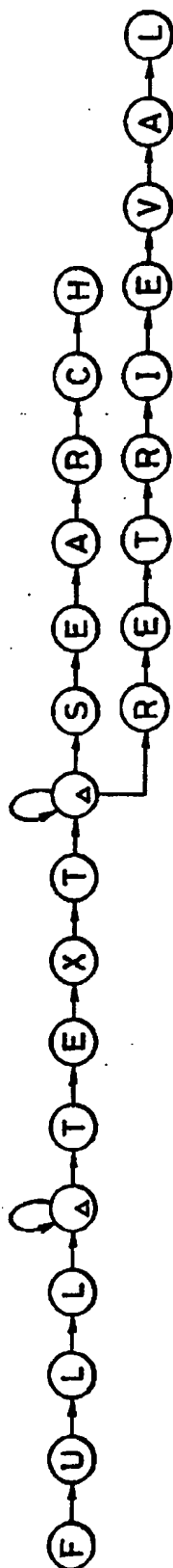


FIG. 3

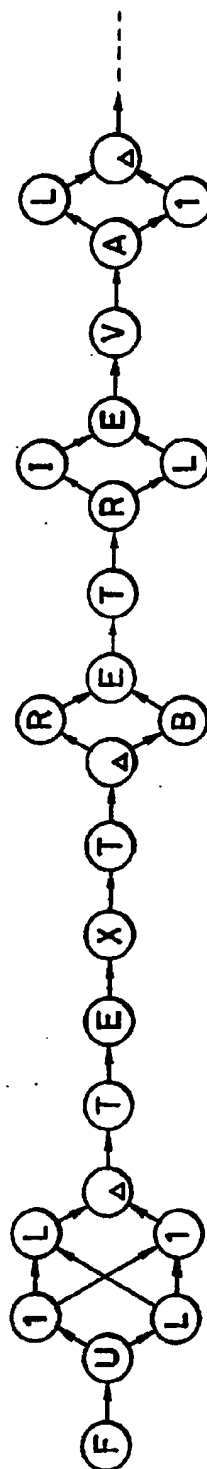


FIG. 4

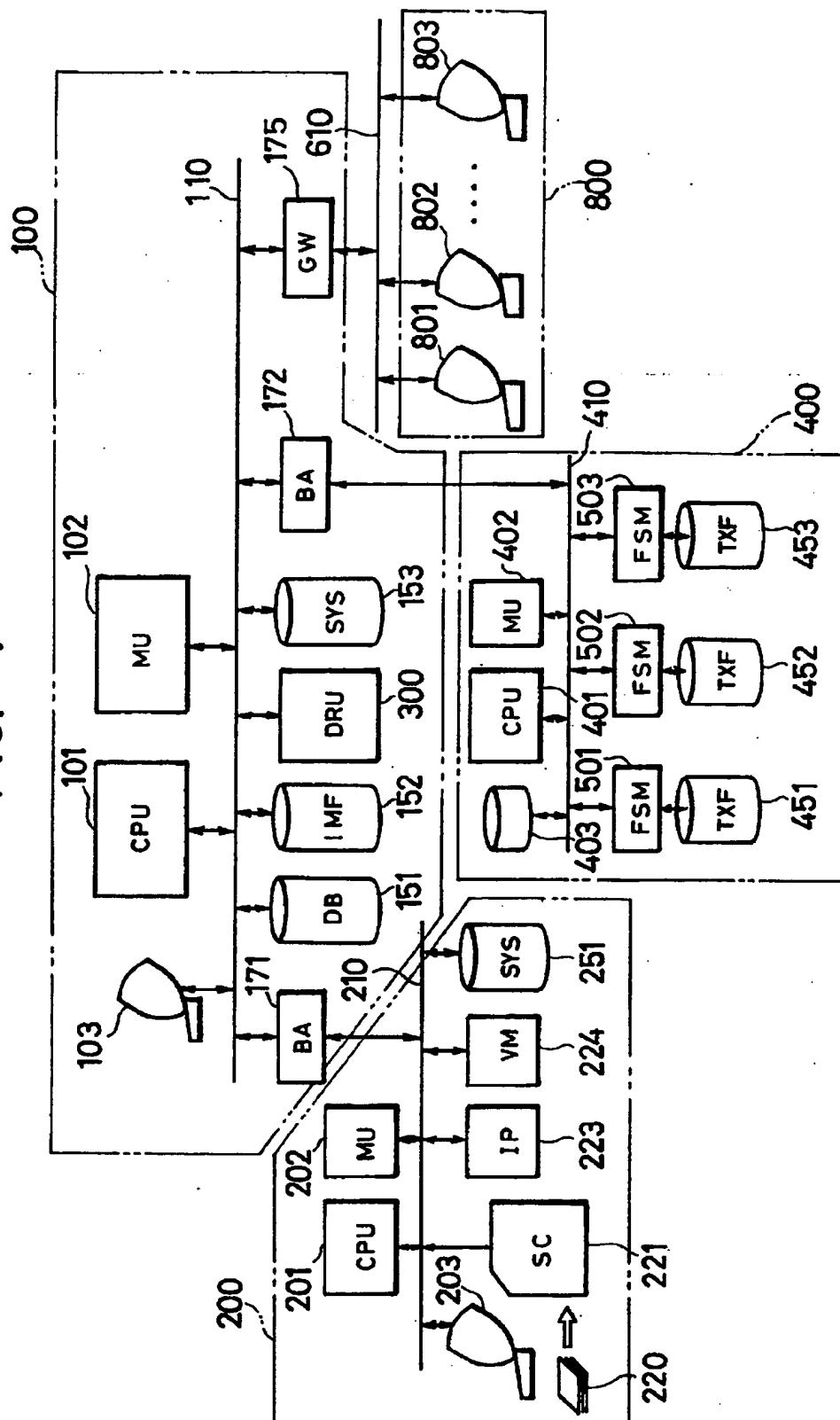


FIG. 5

MAIN-DIR			TITLE	AUTHOR	CLASS	PUBL #	VOL	NO	PP	KWD	ABS	TXT	IMG
DOC #	ID	NP											

FIG. 6(a)

TXT-LOC			
TXTID	VOLSER	UNIT	

FIG. 6(b)

IMG-LOC			
IMGID	VOLSER	UNIT	

FIG. 7

PUBLISHED						
PUB #	TITLE	YEAR	PUBLISHER	ADDR	ASSOC	

FIG. 8

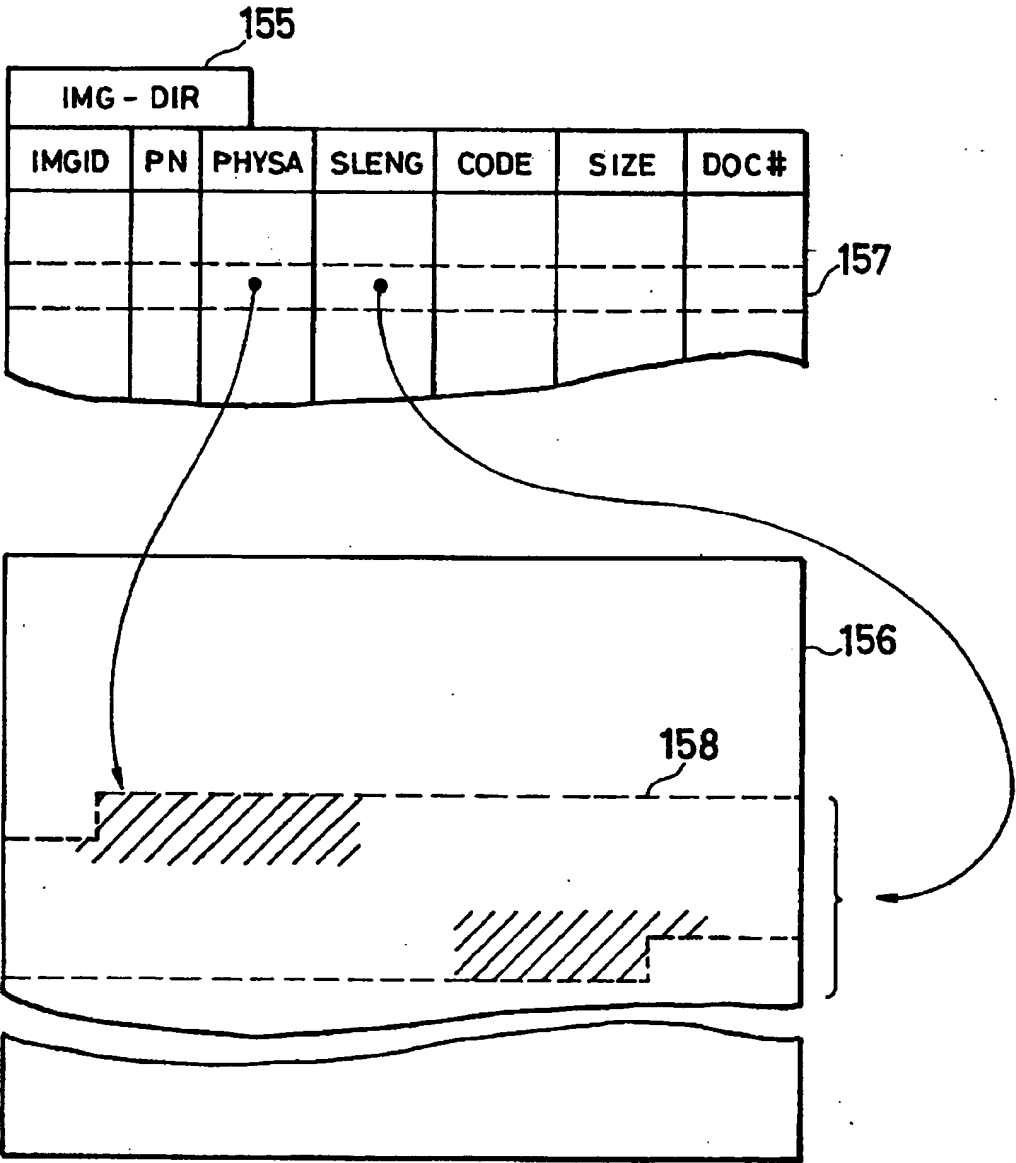


FIG. 9

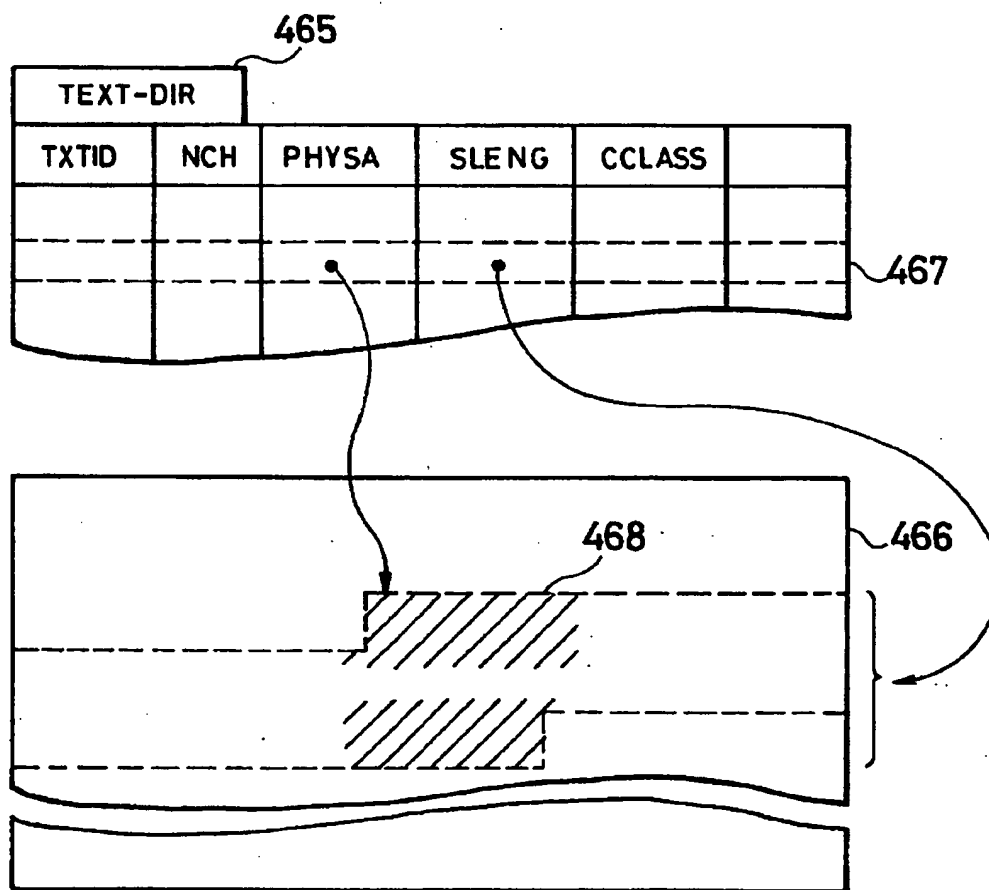


FIG. 10

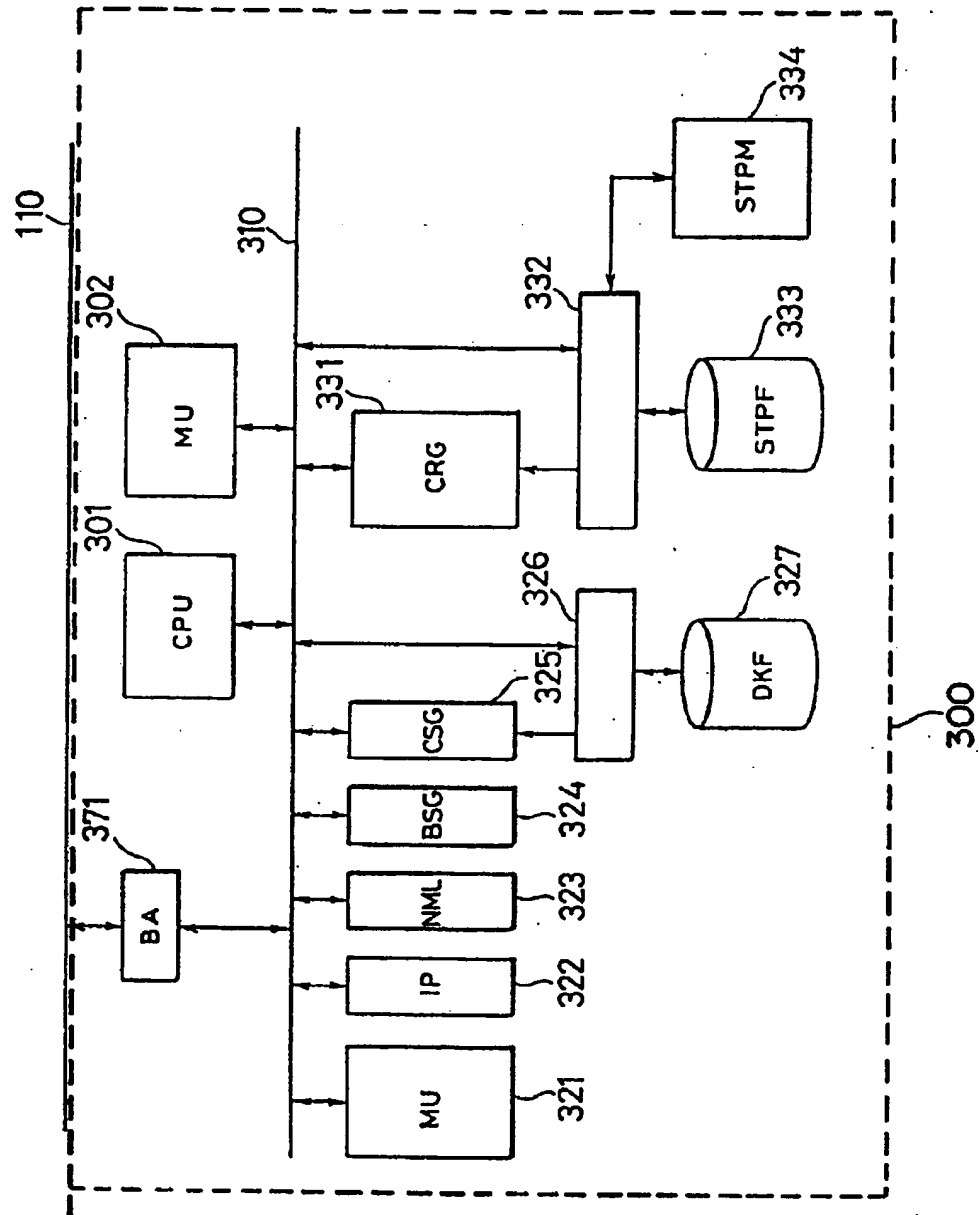


FIG. 11

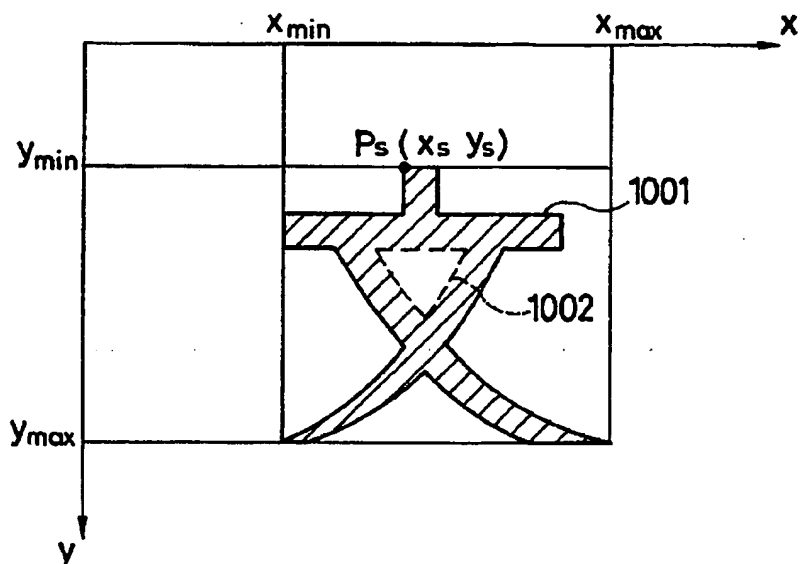


FIG. 12

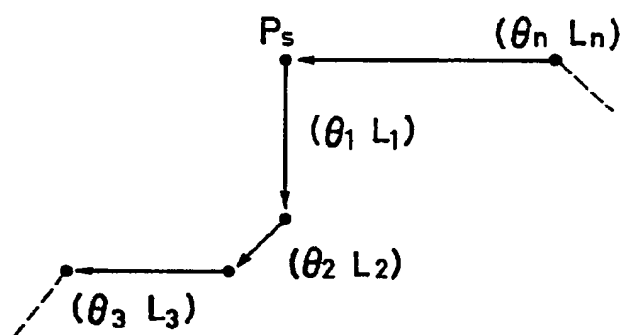


FIG. 13

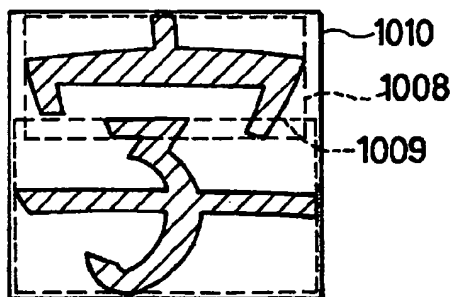


FIG. 14

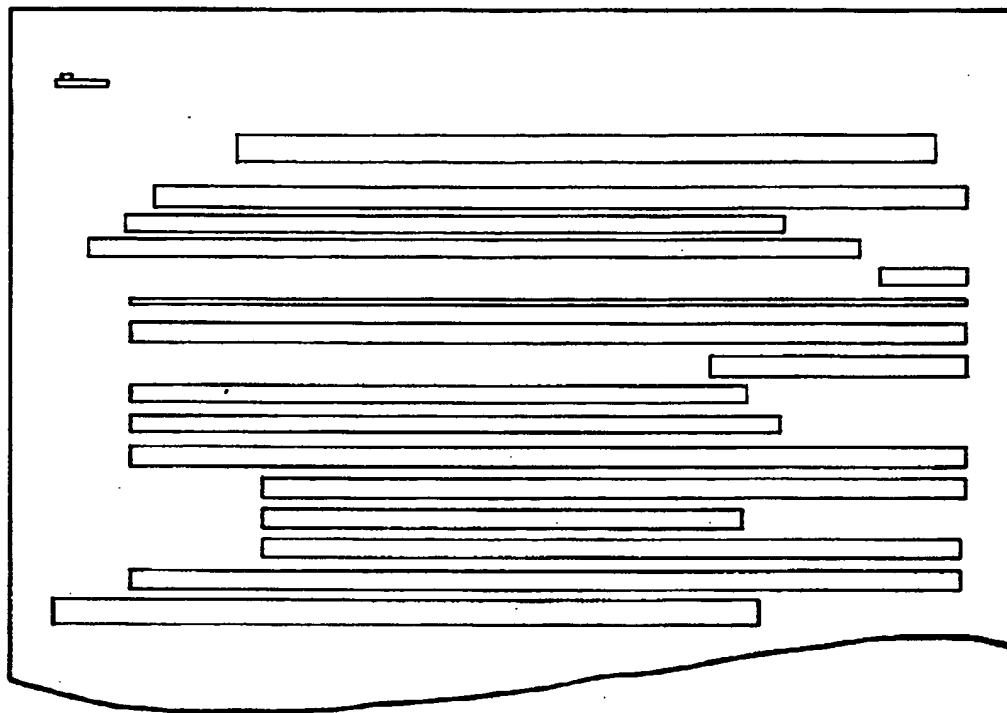


FIG. 15

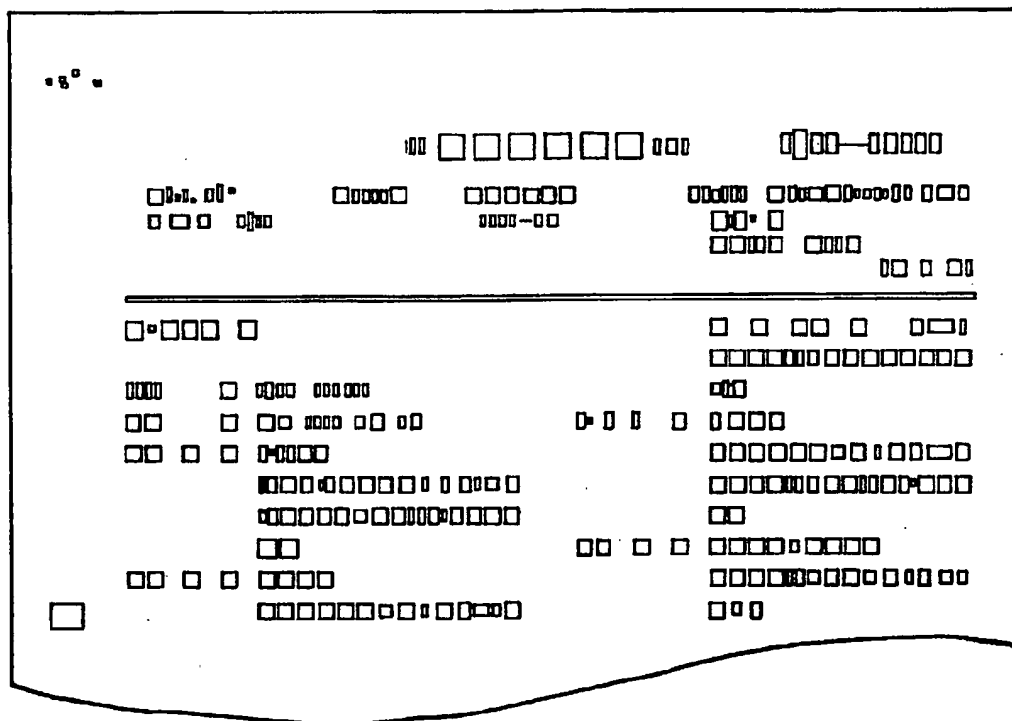


FIG. 16

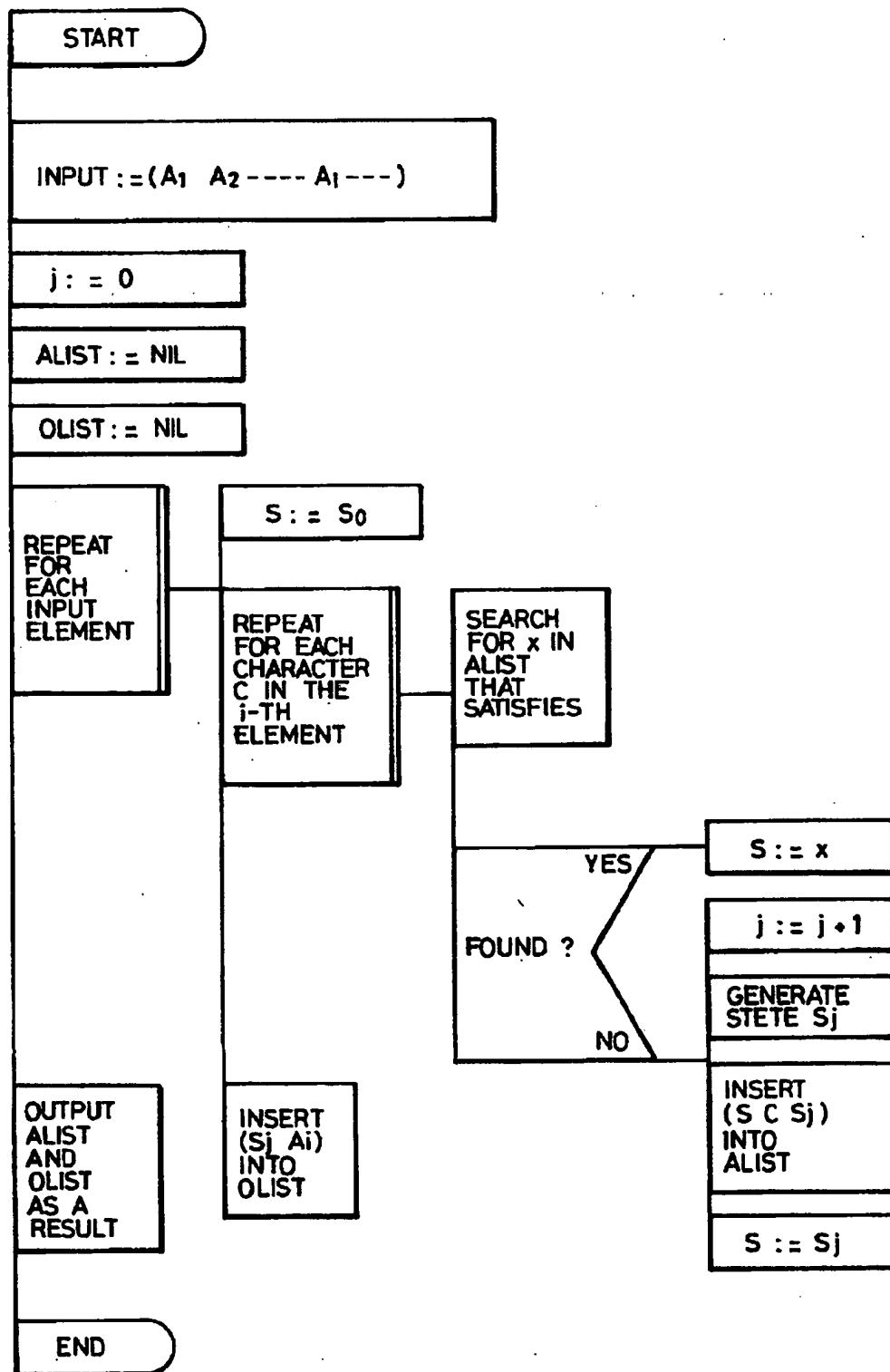


FIG. 17

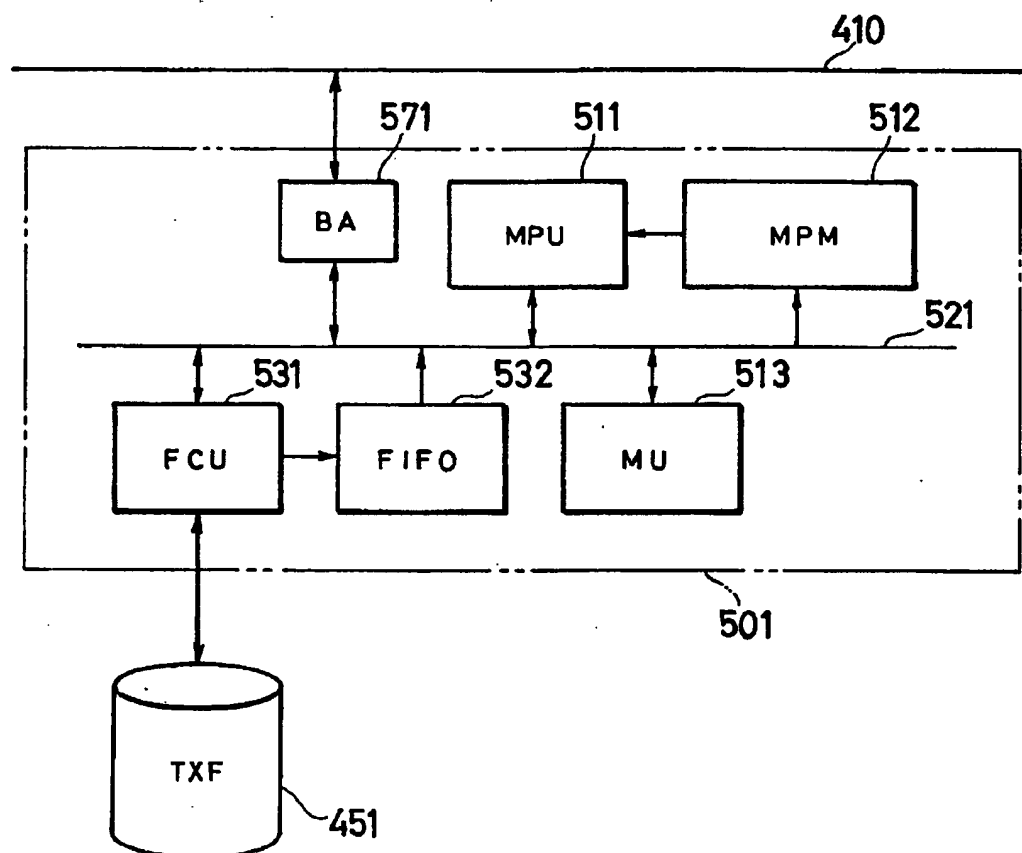


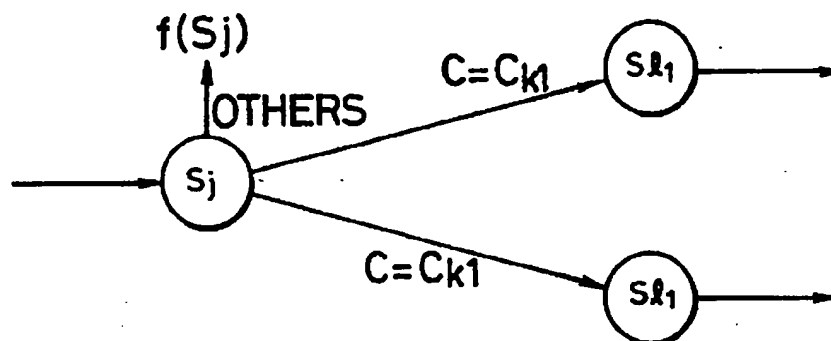
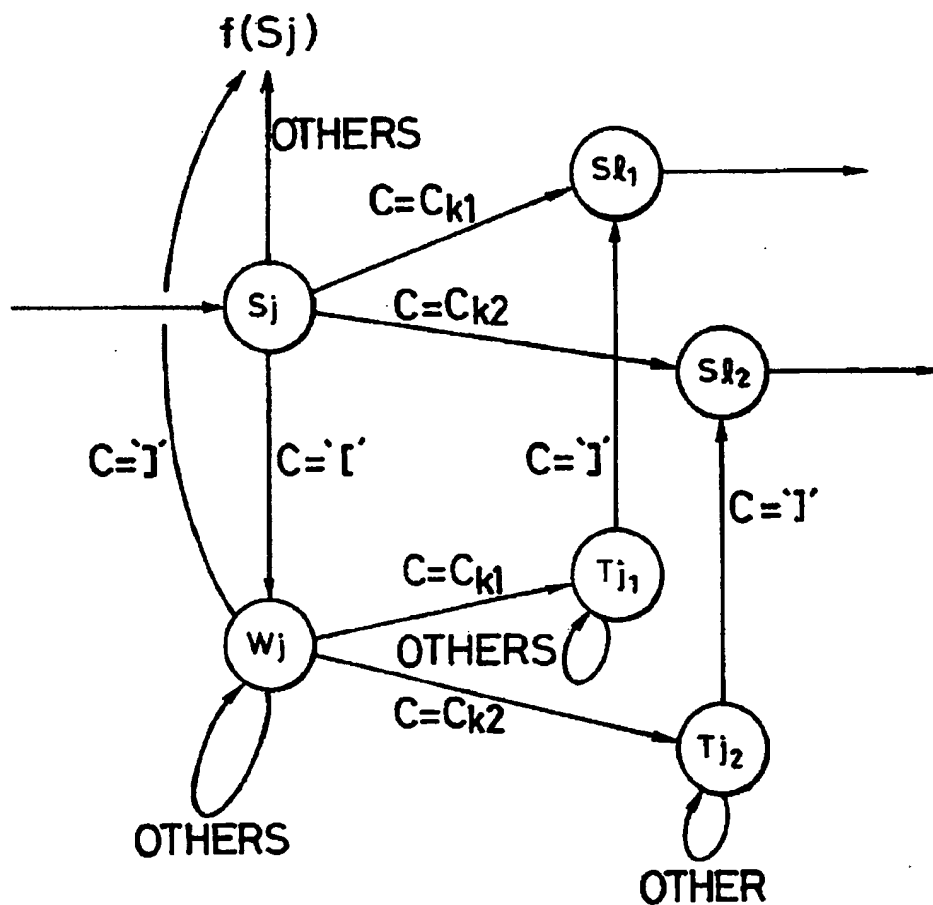
FIG. 18(a)**FIG. 18(b)**

FIG. 19(a)

	----- S _j -----	---- W _j ----	-----T _j -----
C ₁			
C ₂			
C ₃			
⋮			
C _k			
I			
J			

FIG. 19(b)

S	A
S _j	A _i
⋮	⋮

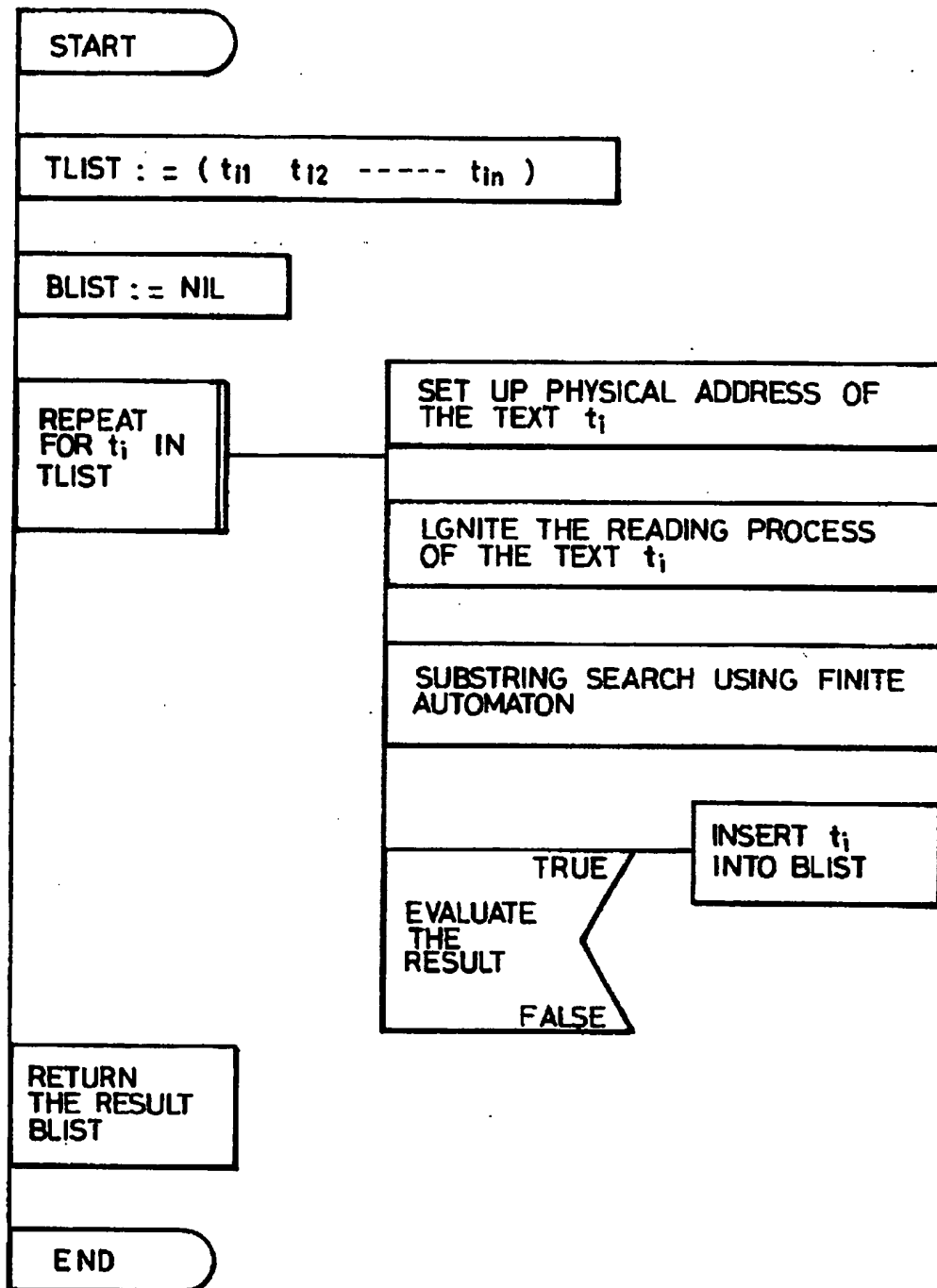
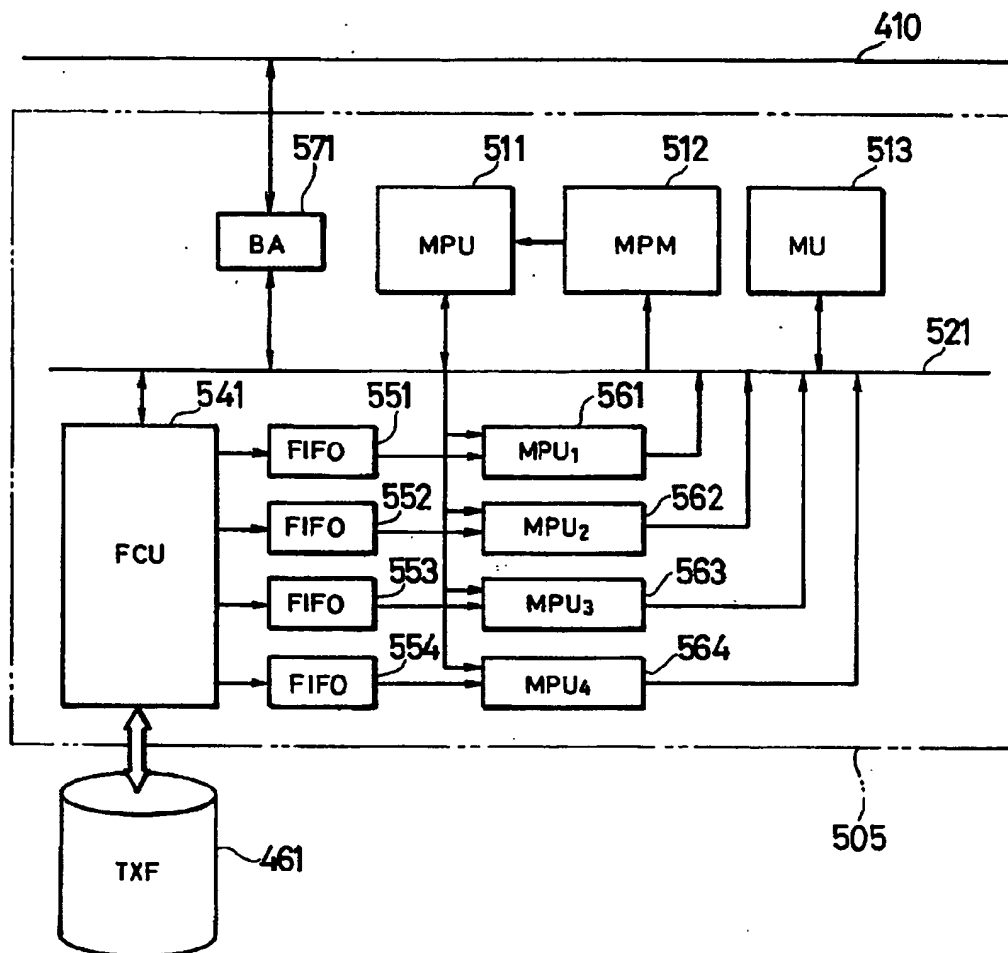
FIG. 20

FIG. 21



DOCUMENT STORAGE AND RETRIEVAL SYSTEM FOR STORING AND RETRIEVING DOCUMENT IMAGE AND FULL TEXT DATA

This is a divisional of application Ser. No. 07/139,781, filed Dec. 30, 1987, now U.S. Pat. No. 5,265,242 issued on Nov. 23, 1993 which is a divisional of parent application Ser. No. 06/894,855, filed Aug. 8, 1986 now abandoned, which was continued as Continuation application Ser. No. 07/559,994 filed Jul. 30, 1990 which issued as U.S. Pat. No. 4,985,863.

BACKGROUND OF THE INVENTION

The present invention relates to a document storage and retrieval system for filing documents as an image, and is particularly concerned with a document storage and retrieval system capable of full text searching.

The typical information retrieval system has hitherto provided a retrieval of data chiefly according to a keyword and a classification code. Bibliographic information and patent information have been processed to form a data base by means of the system mentioned above. Mainly bibliographic information including abstracts in it coverage is processed for a data base here, but the situation is such that only a part of its function is realized to cope with the true need of information retrieval. That is, even if a document or patent conceivably relevant is found, there is the need to search among a lot of bookshelves to obtain the text.

Meanwhile, an optical disk capable of storing a mass data has now been available for loading the text in the data base to provide a so-called original document information service, thus coping with a social need. A paperless documentation at the Patent Office is so planned accordingly. In these systems, volumes of documents are stored in optical disks in the form of image data, and a conventional information retrieval technique based mainly on a keyword search is applied.

However, the conventional information retrieval technique is only effective to orders of tens to hundreds, and hence a further technique for squeezing relevant documents to 1/10 in number or so is desired. One method is that in which an original document (text) stored as image data is called onto a terminal and read visually by a retriever. The method is secure in principle, however, documents amounting to hundreds maximumly are too many to read out in the form of image data, and reading one by one visually is not efficient practically as a matter of course.

On the other hand, the conventional method based on the keyword and classification code must be updated all the time for the classification system itself changes as time passes, thus leaving an intrinsic problem. For example, volumes of documents classified already cannot be modified practically as the classification system is subjected to modification later. Documents and patents recording a progress of science and technology are in content and hence of value because they provide a new data conception which often is not included in the conventional classification system. For this purpose, it is impossible to define beforehand the keyword and the classification system representing a conception originally, which is a problem essentially for the information retrieval system.

For the reason as mentioned above, it is desirable to provide a method which will retrieve contents with reference directly to the text of a document. According to the method for referring to the text, a retrieval can be practiced by means of a vocabulary recognized as a conception which was not

deemed to be important when the document was registered in a data base but is taken new at the point of time of retrieval. Or otherwise, an important document can be searched out directly without a "filter" or an indexer (specialized for giving index) at the time of registration.

To satisfy such a requirement, it is necessary that a character pattern is extracted from the document as image data and the text is replaced by a character code, and a character recognition technique may be applied therefor. However, a document or a printed document, for example, which is an object for filing is not perfect character recognition from the point of view diversification of the kinds of print quality and font. In a conventional optical character reader, imperfect recognitions such as error, rejection and the like are subjected to checks and corrections by operators. (For example, "Introduction to Character Recognition" by Hashimoto, Ohm-Sha, 1982, pp. 153-154) Accordingly, even if the recognition precision is extremely high, a method for checking visually a result obtained through recognizing the text is not realistic where the amount of documents is very large, and hence a document filing system with images as the main constituents which is available for text retrieval has not been realized until now.

SUMMARY OF THE INVENTION

An object of the invention is to provide a document storage and retrieval system having a full text retrieval function with reference directly to the text of a document by solving the problems referred to above.

In order to attain the above-mentioned object, the invention stores and retrieves both the document image data and full-text data, where full-text data is used to support the full-text search capability, and the image data is used to present or display the contents of the retrieved documents to the retriever. This system inputs character strings at the retriever's request, and searches for these strings in character strings in the full-text data. By searching for the corresponding image file identifiers in the image file directory, the locations of the corresponding document image data are identified and the retrieved document images are displayed onto the document retrieval terminal.

This system further recognizes the contents of the documents from the image data and stores the resulting text data to support the full-text retrieval capability. To overcome the problem of insufficient character recognition accuracy, the character recognition module of this system outputs multiple candidates of character codes when more than one characters have very high similarity values, thereby avoiding misrecognition. The full-text data so created therefor includes some ambiguity. For example, an ambiguous text is represented as "... S[mw] [il] th ...", where [mw] represents two characters "m" and "w" which are the candidates for the recognized character, and [il] represents "i" and "l" which are the candidates having the most similarity. The full-text search mechanism of this system can identify that a substring "Smith" is included in the string documents with high accuracy even from the full-text of document recognition results.

As shown in FIG. 1, a document 10 is transformed into a predetermined special character notational expression as indicated by 20 in the system according to the present invention. The symbol string used is that provided in languages such as LISP. It follows a notation called S-expression. A process in which the document (image) is transformed into a notational expression 20 is called document understanding or document recognition. The notational

expression signifies roughly the following. That is, the document is numbered 99, the class is "Technical Paper", VOL=5, NO=7, the author is named "Peter S[mw] [il]th", the title is "Fu[11][11]ΔTextΔ[RB]etr[i]e . . .", the text is ". . . Fu[11][11]ΔTextΔse[ao]rch . . ." and so forth. Here, Δ indicates a blank (space) and so forth.

In the character recognition, that of ambiguity includes, in most cases, a character pattern which can hardly be coped with normally.

For retrieval, meanwhile, a user inputs "FULLΔTEXTARETRIEVAL" from a keyboard. Generally, there are such languages as will express the same meaning in different words, and in this case "FULLΔTEXTΔ SEARCH" has also the same meaning. While handling such ambiguity automatically, the system is capable of searching documents having the same character string.

A plurality of partial character strings to be found out of the sentence to be retrieved are expressed by a finite state automaton as shown in FIG. 2. The title character string which is one of the sentences to be retrieved as exemplified in FIG. 1 can be expressed similarly by the automaton of FIG. 3. In this case, however, there is no distinction between a capital letter and a small letter. The invention provides a text search (character string retrieval) function in case there is present an ambiguity (a plurality of possibilities, or the state wherein elements which cannot be decided identically are present) on both searching key (partial character string) and sentence to be retrieve, which is a third principle.

A method given in a report [by A. V. Aho, et al. "Efficient String Matching: An Aid to Bibliographic Search," Communications of the ACM, Vol. 18, No. 6, 1975] is well known for searching a plurality of partial character strings out of an unambiguous text by the infinite state automaton.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a drawing showing a document image and a result of document understanding;

FIG. 2 is a state transition diagram of a synonymic character string generated from a partial character string;

FIG. 3 is a state transition diagram of a character string as a result of character recognition which includes ambiguity;

FIG. 4 is a system configuration drawing of a first embodiment;

FIG. 5 is a table of the main directory keeping the bibliographic data;

FIG. 6, shows tables for storing location information of text data and image data;

FIG. 7 is a table storing publication information;

FIG. 8 shows an image file directory and its relationship with the body file; and

FIG. 9 shows a text file directory and its relationship with the body file.

FIG. 10 is a block diagram of a document recognizer;

FIG. 11 is an explanatory drawing of a rectangular area surrounding a character pattern;

FIG. 12 is a drawing illustrating a contour expression method for describing a pattern;

FIG. 13 is a drawing illustrating a relation between pattern components and character pattern;

FIG. 14 and FIG. 15 are drawings showing a result of segmenting rows and columns respectively by means of a bottom-up segmenter;

FIG. 16 is an explanatory drawing of an algorithm for obtaining a state transition list from a character string aggregation;

FIG. 17 is a block diagram of a flexible string matching circuit;

FIG. 18 is an extended finite state automaton permitting an ambiguous character string;

FIG. 19 is a state transition table of the extended finite state automaton;

FIG. 20 is a drawing illustrating a program of FSM circuit;

FIG. 21 is a configuration drawing of a flexible string matching circuit in a second embodiment.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

The invention will now be described with reference to illustrative examples. FIG. 4 is a configuration drawing of a document storage and retrieval system forming one embodiment of the invention. The system comprises a control subsystem 100 providing a general control and a data base function, an input subsystem 200 for inputting a document and others and registering in a file, a document recognizer 300 for recognizing documents, a text search subsystem 400 for carrying out a high-speed text search, and a terminal subsystem 800 for carrying out a retrieval.

A configuration and a flow of operation of each subsystem will be described in detail below.

The input subsystem 200 has a central processing unit (CPU) 201 for controlling the subsystem, a main memory 202, a system file 251 and a terminal 203 as a basic division. The subsystem is controlled by operation from the terminal 203, an image on each page of a document 220 is read optically by a scanner 221, and digitized image data is stored first in a video memory 224 by way of a bus 210. The image data is then subjected to a redundant compression on an image processor (IP) 223, transformed into MH (Modified Huffman) code or MR (Modified Read) code and then returned to another area of the video memory 224.

The inputted document image is displayed on the terminal 203 for confirmation, and the operator is capable of inputting bibliographical items such as the title, author's name, creation data and others while observing the image displayed thereon. As will be described hereinlater, bibliographical items of a formatted document can be read automatically through document understanding, however, bibliographical items of a not-formatted document and items of information which are not entered in paper must be inputted manually. For example, it is natural that a classification code of document contents defined by users and a keyword which is not present on paper should be inputted by the operator. Then, a value and position setting of each document must be arranged independently by a user of the document, which can be inputted from the terminal 203. A data of such bibliographical items and others inputted as above is correlated with an image data (compressed data) in the video memory 224 and is then loaded in the main memory 202.

Here, each document is given a proper number (document ID) and stored in the memory so as to draw image data and bibliographical items using the proper number of the document as a key. The document proper number can be expressed, for example, by coupling an identifier number ("INSYS 01" and the like) of the subsystem to the character string indicating date and time. For example, INSYS01.850501.132437 indicates a document inputted from an input subsystem INSYS01 at 13 h: 24 m: 37 s on May 1, 1985. There may be a case where the input time is important according to application of the system, and hence it functions as a time stamp otherwise.

Now, whenever a predetermined quantity of the document is accumulated in the subsystem 200 or a predetermined command arrives from the terminal 203, an interrupt signal is sent to a bus adapter 171.

A control subsystem 100, sensing the interrupt signal, reads a predetermined address in the memory 202 of the input subsystem 200. The contents of a request of the input subsystem can thus be decided.

An operation follows as described below upon request of a registration of the inputted document in a data base.

The central processing unit (CPU) 101 is acquainted with the proper number of documents stored temporarily in the input subsystem 200 according to a predetermined program in a main memory 102 and further with a memory address of bibliographical data (bibliographical items) relating thereto and image data.

The control subsystem 100 has a data base file 151 for storing and managing symbolic data such as bibliographical data and the like, and an image file 152 for storing and managing the image data.

The bibliographical data read out of the input subsystem 200 is written as a new record in a data base (loaded in the file 151) which is given in the form of the table in FIG. 5. The table of FIG. 5 is named MAIN-DIR (main directory) and has the following data columns.

DOC#: A serial number of document registered in the system
 ID: A document proper number given by the input subsystem.
 NP: A page number constituting the document.
 TITLE: A title (character string)
 AUTHOR: An author's name (permitting iteration of plural data).
 CLASS: A symbol indicating classification, kind and the like of documents.

PUBL#: A number of publication registered in the system (detail being managed on the table shown in FIG. 7.)
 VOL, NO, PP: Volume, number, page.
 KWD: A plurality of keywords.
 ABS: A text proper number of abstract expressed as a character code string (text data).
 TXT: A text proper number as a character code string.
 IMG: A proper number of image data. Since the image data is managed at every page, a plurality of image proper numbers are recorded.

In registration of the bibliographical data, only such data of the above columns as will relate partly to the bibliographical data is written newly.

Next, the image on a page constituting each document is read to the control subsystem 100 from a predetermined storage area of the input subsystem and is then stored sequentially in an empty area of the image file 152. Each image (page unit) is concurrently given an image proper number (IMGID). Then, a volume number (VOLSER) of the file having loaded the image data therein, a file unit number (UNIT), a loading physical address (PHYSA) in the file, a record length (SLENG) in the file and others are written in tables shown in FIG. 6(a) and FIG. 8. The image proper number IMGID given newly is also recorded in IMG column of the table MAIN-DIR (FIG. 5).

Here, a table IMG-LOC shown in FIG. 6(b) is particularly effective when the image file 152 is constituted of a plurality of driving devices or a plurality of volumes, managing the location of each image. As a matter of course, it is updated

at every operations for demounting and mounting the volume by operators.

Then, FIG. 8 shows a directory provided at each volume of the image file 152, and the following columns are provided therein.

IMGID: An image proper number.
 PN: A serial page number (1 to n) in a document.
 PHYSA: A physical address in a volume.
 SLENG: A record length (sector number, for example).
 CODE: An image compression code name.
 SIZE: An image size (pixel number).
 DOC#: A document serial number.

Then in the drawing, data in the column PHYSA of a record 157 indicates a leading address of image data 158 in an image data area 156 in the image file.

Now, whenever the above operations come to end, the system is ready for retrieving the bibliographical items and the keyword from the terminal group 800.

A retrieval condition inputted from the retrieving terminal is transmitted to the CPU 101 of the control subsystem 100 by way of a gateway 175. A retrieval of a table MAIN-DIR (FIG. 5) in the data base file 151 is carried out according to a predetermined retrieving program of the memory 102. It goes without saying that indexing (for high-speed retrieval such as hashing, inverted file and the like) is applied to main columns of the table MAIN-DIR.

As a result of retrieving, a list of DOC# from the table MAIN-DIR (FIG. 5) and a list of image proper number IMGID are made out and stored in a predetermined area of the memory 102. Upon request for display from the retrieving terminal, a position in the image file is identified by means of a table IMG-LOC 154 (FIG. 6(b)) and a table IMG-DIR 155 (FIG. 8), and the image data is read successively onto the memory 102. The image data thus read out is transmitted to the retrieving terminal in turn and then displayed on a screen according to an indication on the terminal.

A managing method for the text used for full text retrieval will be described, next.

As described in the main directory MAIN-DIR (FIG. 5), each document is stored and managed not only for image data but also for text expressed in a character code string. In the example, the abstract and the text are stored and managed in text files 451, 452, 453 as a text. Each text (character string) is given a proper text number and recorded in columns ABS and TXT of the table MAIN-DIR (FIG. 5), a column TXTID of the table TXT-LOC shown in FIG. 6(a), and a column TXTID of the table TEXT-DIR shown in FIG. 9.

FIG. 9 indicates a method for storing and managing texts in the text files 451, 452, 453. In the drawing, a text body is stored one-dimensionally in a file storage area 466. Each text (one character string) is given a proper number TXTID and managed in a directory table TEXT-DIR 465.

TXTID: A text proper number.
 NCH: A total number of characters constituting the text.
 PHYSA: A physical address in which the text is recorded.
 SLENG: A record length on a storage medium of the text.
 CCLASS: A class of characters expressing the text (Chinese character-mixed Japanese statement, English statement, Roman character, kana character and others).

A record 467 of the table 465 indicates that the text expressed by the record is a portion 468 in the storage area in the file.

On the other hand, as shown in FIG. 4, the text can be recorded in a plurality of volumes, and the text directory is that of managing the text in each volume. When the plural volumes are mounted, it is necessary that a presence of a text in any of the volumes be known, and the table TXT-LOC shown in FIG. 6(a) manages the location of each text. A volume serial number VOLSER in which the text having the text proper number TXTID is recorded, and a file unit number UNIT in which the volume is mounted is managed. TXT-LOC will be updated automatically as a matter of course when a physical volume is demounted or newly mounted by operators.

Then, when input of document images, input of bibliographic items and registration of documents are constituting a large operation is completed, a text recognition (document understanding) of the registered document is carried out by the document recognition apparatus 300. An input of the recognition apparatus is the document image 10 shown in FIG. 1 in an image file 152, and a recognition result output is a notational expression 20 shown likewise in the drawing. A text portion of the abstract and the text in the notational expression 20 is stored newly and so managed by he text files 451 to 453 as described hereinabove.

The document recognition will be described with reference to a detailed block diagram of the document recognition apparatus shown in FIG. 10.

The recognition apparatus 300 is connected to a bus 110 of the control subsystem 100 through a bus adapter 371 and controlled by CPU 301. A memory 302 stores data of a program and a parameter for controlling operation of the apparatus.

An image data to be recognized is transmitted from the image file 152 to a memory 321. The image data is coded through compression, decoded to a bit expression image by an image processing circuit IP 322 and is again stored in the memory 321. Then consecutively, a contour extraction of the pattern is carried out by the IP 322 from the image decoded to a bit expression, and a result of extraction is again loaded in the memory 321.

The extracted contour data is expressed as follows:

$$(i \text{ Ci } x_{\min}, x_{\max}, y_{\min}, y_{\max}, \theta_1, \theta_2, \dots, \theta_n) \quad (1)$$

where i represents a contour proper number (1, 2, 3, ...), and Ci represents a class of the contour. Then, $Ci=0$ represents an outer contour (a full line 1001 in FIG. 11), and $Ci=1$ represents an inner contour (a broken line 1002 in FIG. 11). Those $x_{\min}, x_{\max}, y_{\min}, y_{\max}$ represent a coordinate of the vertex of an outer quadrangle of the contour, each, as shown in FIG. 11. Further, (x_p, y_p) is a coordinate of one point P_s of the contour length (or, for example, the point found first by contour retrieval). With the point P_s as an origin, as shown in FIG. 12, the contour data itself is expressed by rows of sets of a quantized direction code θ and a pixel number L with the same direction continuing therefor.

Next, an inclination correction circuit 323 detects a tilt angle arising at the time of document input from the contour data given by the expression (1), corrects the contour data accordingly and then rewrite it to the memory 321. For example, a system disclosed by the inventor in Japanese Patent Application No. 152210/1985 may be employed for the inclination correction algorithm.

From a portion of the contour data corrected for inclination ($x_{\min}, x_{\max}, y_{\min}, y_{\max}$), a raw segmentation and a column segmentation are carried out on a bottom-up segmenter (BSG) 324.

The bottom-up segmenter BSG inputs the data expressed in the form of expression (1), generates a pattern list given by the expression (2) and loads it in the memory 321.

$$(j \text{ } x_{\min}, x_{\max}, y_{\min}, y_{\max}, j)$$

(2)

Here, j represents a pattern proper number, the pattern is defined as a rectangular area not overlapping mutually, and the expression (2) further defines vertex coordinates of the rectangular area. For example, rectangular areas 1008, 1009 indicated by broken lines in FIG. 13 are inputs of the BSG, however, a rectangle 1010 is obtainable through the BSG. The rectangles 1008, 1009 are made of one contour each to be an element, and the rectangle 1010 is a pattern forming one character. An element constituting the pattern j is obtainable through searching the rectangle included in a rectangular area defined by the expression (2) from the contour data of the expression (1). It can be obtained separately and loaded as data. A result of row segmentation and another result of column segmentation are shown diagrammatically in FIG. 14 and FIG. 15 respectively.

A character segmentation division (CSG) 325 extracts the pattern constituting a character from the above pattern list with reference to a document knowledge arranging regulations such as document form and the like. As shown in FIG. 10, the document knowledge is loaded in a document knowledge file (DKF) 327.

Structural regulations of the layout of such as a title, author's name, author's belonging, abstract, text and the like are stored according to each kind of documents in the document knowledge file together with a parametric knowledge such as the size of font. The knowledge is described in a format description language. The language disclosed in Japanese Patent Application No. 122424/1985 may be used as a format description language.

The character segmentation division CSG operates for integration of a pattern constituting one character which has been divided into two patterns or more or, to the contrary, for compulsory separation of two or more characters which has been fused through contact into one pattern.

The character segmentation division CSG outputs the number of the patterns constituting each character in a list for each item such as the title, abstract or text as the result of processing. For example:

$$(\text{ABSTRACT } "j_1 j_2 \dots [j_n j_{n+1} j_{n+2}] \dots j_m") \quad (3)$$

represents that the abstract is constituted of a string of characters expressed by a pattern number j_k . Here, $[j_n j_{n+1} j_{n+2}]$ represents that the character in a combination of three patterns j_n, j_{n+1}, j_{n+2} .

A character recognition division (CRG) 331 extracts the contour data constituting each character pattern, as described hereinabove, from the above-mentioned pattern list (expression (3), for example) and the contour data (given by expression (1)) on the memory 321, and transforms it into a data structure ready for feature extraction.

Since a known art may be employed as the character recognition technique, a detailed description will be omitted here, however, after a feature is extracted from the contour data, each character can be recognized through a pattern matching with the standard pattern in a standard pattern file 333. In FIG. 10, a memory STPM 334 is one for storing a standard pattern with high reference frequency, aiming at a high-speed processing.

The result of the character recognition is output, as described hereinabove, by the notational expression 20 shown in FIG. 1. In the process of final decision in the character recognition, when a similarity obtained as a result of pattern matching satisfies an expression (4), a character category (character code) ω_k for giving the similarity is output.

9

$$p_k \geq p_l \min(p_k - p_l) \geq \epsilon k \text{ for } l=1, 2, \dots, K \quad (4)$$

where p_k is a similarity to the character category K , K is a total category number, and ϵ is a relative threshold.

If the expression (4) is not satisfied, then an aggregation of the character category $\{\omega_k | k=k_1, k_2, \dots\}$ satisfying an expression (5) is output within two special character codes. For example, a character (code) string $\omega, \omega_{k_1} \omega_{k_2} \dots \omega_e$ is output. Here, ω represents "[", and ω_e represents "]"

$$\begin{aligned} p_k &\geq p_l \text{ for } l=1, 2, \dots, K \\ p_k - p_{k_1} &\leq \epsilon_1 \\ k_1 &\{1, 2, 3, \dots, K\} \end{aligned} \quad (5)$$

In case a similar character is present and the expression (4) is not satisfied by the above processing, a recognition result "FU[L1][L1]TEXTASEA[RB]CH" is obtainable, for example, in response to the input pattern "FULLTEXTSEARCH". The recognition result is buffered on the memory 321 and then transmitted to the memory 102 (FIG. 4) collectively.

In the control subsystem 100, a maximum text proper number is detected with reference to the table TXT-LOC (FIG. 6), and a character code string (text) of the recognition result is registered with a value added by 1 as a new text proper number. The registration is carried out with respect to the main directory 153, the table TXT-LOC and the table 465 (FIG. 9), and the text data itself is loaded in any of the text files 451 to 453.

Now, the document to which a text data is given as above is ready for retrieving using the text search subsystem 400.

Next, the text search subsystem 400 for retrieving text contents and its operation will be described in detail.

A request for text content retrieval or ABS="TEXTARETREIVAL", for example, which is so made from the terminal 800 is transmitted first to the control subsystem 100. In the subsystem 100, where the document to be retrieved has already been narrowed down through keyword retrieval or other means, a proper number of the text incidental to the document is selected from the main directory MAIN-DIR 153, and an expression (6) for the list of proper numbers of the texts to be retrieved is made out according to each text file with further reference to the table TXT-LOC.

$$\begin{aligned} (u_i v_i (t_{i1} t_{i2} \dots t_{iM})) \\ i=1, 2, \dots, M \end{aligned} \quad (6)$$

where u_i is an i -th file unit number, v_i is a volume serial number, t_{ik} is a k -th text proper number to be retrieved on the volume. Then, M is a maximum number of the text file unit.

On the other hand, when the document to be retrieved has not been narrowed, a special symbol (expression (7), for example) is sent to the whole text file.

$$(u_i v_i^*) i=1, 2, \dots, M \quad (7)$$

The expression (6) or (7) and the partial character string ("TEXTARETREIVAL", for example) are transmitted to a memory 402 of the text search subsystem 400 by way of a bus adapter 172.

In the subsystem 400 (FIG. 4), a hetero-notation generation processing and a synonym processing of the transmitted partial character string are carried out according to a predetermined program in the memory 402 the term "hetero-notation" refers to words having the same meaning but a

10

variation in form or spelling, such as "center" and "centre" or "DATA" and "data". A hetero-notation generation convention and a thesaurus are stored in a file 403.

"TEXT SEARCH" will further be obtainable through referring to the thesaurus. Further, with reference to the hetero-notation generation, a method disclosed in Japanese Patent Application No. 150176/1985 may also be employed.

As the result of the above-mentioned processing, an aggregation of character strings ("TEXTARETREIVAL" "TEXTASEARCH") is obtainable, after all, to "TEXTARETREIVAL". This is indicated by an expression (8).

$$(A_1 \dots A_i \dots A_n) = ("a_{11} a_{12} \dots a_{1m_1}" \quad (8)$$

$$\begin{aligned} &\vdots \\ &"a_{i1} a_{i2} \dots a_{im_i}" \\ &\vdots \\ &"a_{n1} a_{n2} \dots a_{nm_n}" \end{aligned}$$

where n is a number of character strings, m_i is the length of an i -th character string, a_{ij} is a character code j -th from the lead of an i -th character string A_i .

The subsystem 400 further transforms the expression (8) representing the character string aggregation into a state transition list (9) representing the finite automaton illustrated in FIG. 2 according to a predetermined program.

$$a \text{ list} = ((S_{j1} \quad C_{k1} \quad S_{n1}) \quad (9)$$

$$\begin{aligned} &\vdots \\ &(S_{j2} \quad C_{k2} \quad S_{n2}) \\ &\vdots \\ &(S_{jm} \quad C_{km} \quad S_{nm}) \end{aligned}$$

where each element of the list a list (9) implies that when the character C_{ki} is inputted (or coincides therewith) in the state S_{ji} , the state can be transmitted to the state S_{ni} . Then in the expression, those which are equal to each other are included in $\{S_{j1}, \dots, S_{jm}, \dots, S_{jn}\}$.

Further, an output list (10) expression is generated.

$$O \text{ list} = ((S_{j1} \quad A_{i1}) \quad (10)$$

$$\begin{aligned} &\vdots \\ &(S_{jp} \quad A_{ip}) \\ &\vdots \\ &(S_{jn} \quad A_{in}) \end{aligned}$$

where $(S_{jp} \quad A_{ip})$ implies that the character string A_{ip} is found at the point of time when reaching the state S_{jp} .

FIG. 16 shows a PAD (program analysis diagram) of the algorithm for deriving the state transition list (9) and the output list (10) from the character string aggregation (8) expression.

Next, a failure transition list (11) expression is obtained from the state transition list (9).

$$f \text{ list} = ((S_o S_{jo}) \dots (S_m S_{jm})) \quad (11)$$

The element $(S_m S_{jm})$ of f list specifies transition of the character C_k inputted in the state S_m to the state S_{jm} with

reference to f list when the state to be transmitted is not specified in a list (9) expression. It may be called generally a failure function.

The f list is provided so as to cope with the case where a reinitialization of the state to S_0 is generally not correct when a matching is successful halfway of a character string but the next character does not coincide in the partial character string matching, i.e., a destination of the predetermined state transition is not found. For example, a retrieval of two partial character strings "CHARACTERARECOGNITION" and "OPTICALCHARACTERAREADER" is assumed. Supposing a sentence reading "OPTICALCHARACTERARECOGNITION" is inputted, a portion up to "OPTICALCHARACTERARE" coincides with the second partial character string but the next character "C" is not for matching. Here, if the state is returned to S_0 to resetting, the automaton processes the ensuing sentence "COGNITION" as input characters, therefore the partial character string "CHARACTERARECOGNITION" will be overlooked after all. Accordingly, the state to be transmitted in the case of failure matching is not S_0 , but the state must stand as matching a transition pass "CHARACTERARE" of the first character string "CHARACTERARECOGNITION".

Then next, the subsystem 400 transmits the state transition list, a list, the output list, o list, and the failure transition list, f list, made out as described above to lower flexible string matching circuits FSMs 501 to 503.

A further detailed block diagram of the flexible string matching circuit 501 is shown in FIG. 17. (The block diagram applies likewise to FSMs 502, 503.)

The above-described three lists, a list, o list and f list, are loaded in predetermined areas of a memory 513 by way of an adapter 571. A microprocessor 511 generates an extended finite automaton shown in FIG. 18(b) in the form of a state transition matrix on the above information according to a predetermined microprogram.

The extended finite automaton that the lists, a list and f list, directly imply has a simple form as shown in FIG. 18(a). The drawing illustrates two transitions

$$\begin{aligned} (S_{C_{k1}} S_{h1}) \\ (S_{C_{k2}} S_{h2}) \end{aligned} \quad (12)$$

in the a list.

The microprocessor 511 extends and transforms the extended finite automaton shown in FIG. 18(a) to the one as shown in FIG. 18(b). The transformation is determined identically. A predetermined partial character string can be searched from the ambiguous text to be retrieved according to the transformation. Here, in the drawing, $f(S_j)$ is a failure function made out of the failure transition list f list, indicating a state of the destination of transition when failing in matching at the state S_j . Then, the state W_j corresponds one-to-one to the state S_j , scanning the ambiguous character string (given within symbols [to]). Further, the states T_{j1}, T_{j2} are states coming out of the state W_j correspondingly to a transition from the state S_j , indicating that the character being retrieved (C_{k1} or C_{k2} in the drawing) has been found in the ambiguous character string.

Practically, the microprocessor 511 is capable of generating the state transition table shown in FIG. 19(a) directly from the two lists, a list and f list. A column (vertical) in the state transition table indicates a current state, and a row (lateral) corresponds to a character (code) inputted under the state. The state to transit next is written in the table. Since

the algorithm for generating the state transition table will be analogized easily from illustration according to FIG. 18, a further description is omitted.

The microprocessor 511 further transforms the output list, o list, into the form of an output table shown in FIG. 19(b) and records it in a predetermined area of the memory 513 together with the state transition table.

A string search algorithm using the finite state automaton is given as below.

```
[String Search Algorithm]
begin
   $\gamma := \text{'false'}$ ;
   $S := S_0$ ;
  While not eof do
  begin
    read (c);
     $S := \text{next}(c, S)$ ;
    if out (S) <> nil
      then  $\gamma := \text{'true'}$ ;
  end;
end;
```

Here, the function next (c, S) is one for obtaining the next state from the state transition table shown in FIG. 19(a) on the character c and the current state S. Further, the function out (S) is one for deciding whether or not an output is present on the state S with reference to the output table shown in FIG. 19(b).

Then, the state is assigned to a unit of one character code in the above description, however, in case the one character code is 2 bytes like Japanese, it is divided into 1 byte each and then the above-described method can be applied thereto.

Next, the text search subsystem 400 accepts the lists (6) expression and (7) expression of the proper numbers of texts to be retrieved, and transmits them to the corresponding FSM as text proper number lists to be retrieved at each FSM. Accordingly, if there exists an object to search in the corresponding text file, each FSM obtains the proper number list ($t_{11} t_{12} t_{13} \dots t_{1n}$). The text proper number list is loaded in the memory 513 (FIG. 17). The microprocessor MPU 511 detects a physical address of each text according to a predetermined program (FIG. 20) in a microprogram memory 512. The text proper number and the physical address are managed by TEXT-DIR illustrated in FIG. 9, and the table can be read out of the file 451 and thus detected.

The microprocessor 511 then reads each text data out of the file 451. A file control division 531 inputs text data (character string) thus read out successively to an FIFO (first-in-first-out) circuit 532. The microprocessor MPU 511 reads characters one by one out of FIFO 532 and verifies whether or not a predetermined partial character string is present according to the finite automaton (FIG. 18(b)) defined in the memory 513. A string matching result b list (FIG. 20) is returned to the memory 402 of the upper processor.

CPU 1 arranges text proper number lists with retrieval conditions coincident with each other which are sent back from a plurality of lower FSM's into one according to a predetermined program and transmits them further to the memory 102 in the upper control subsystem. A document proper number DOC# with partial character strings matched therefor and a proper number of a document image IMGID or a title TITLE can be identified from the text proper number by referring to the main directory 153 (FIG. 5).

The retrieval results are sent back to the terminal 800. Users are capable of calling the image of a desired document to a CRT to display thereon while observing the title and others on the CRT.

A second illustrative example will be described, next. In the example, a configuration of the flexible string matching circuit 501 only is different. FIG. 21 is a configuration drawing of the flexible string matching circuit FSM in the second example.

In the drawing, a secondary storage unit (text file) 461 has a plurality of heads capable of reading a signal simultaneously, and in the example, data can be read out of the four heads simultaneously. The data is transmitted to four FIFO circuits 551 to 554 each by way of a file control unit FCU 541.

On the other hand, retrieval conditions sent from the upper subsystem 400 are interpreted by the microprocessor 551 and then transmitted to microprocessor units MPU₁ 561 to MPU₄ 564 including data memories.

Text data read out of the text file 461 are read to the microprocessor units 561 to 564 each by way of FIFO circuits 551 to 554. The microprocessor units search in parallel a predetermined partial character string from among four character strings (text data) and sends the result back to the microprocessor 511 by way of a data bus 521.

Since the other portions are equal to those of the first example, a description will be omitted.

A third illustrative example will be then taken up for description. In the example, the hardware configuration is the same as those of the first and second examples, but the text searching is different.

In taking up the case where a document to be retrieved is narrowed down by means of a keyword or classification code according to a hierarchical retrieval method, the document screened in the process is generally unevenly distributed to a volume of the text file.

In the system of the example, a text data is stored redundantly in a plurality of text file volumes for multiplicity. According to a predetermined program, CPU 401 (FIG. 4) selects a volume to access so as to even the frequency of access to a plurality of volumes for the texts stored redundantly in the volumes. According to the system, all the flexible string matching circuits operate efficiently, and a high-speed retrieval can be realized as a whole.

In the above example, a multiplicity of the flexible string search circuit is 3 or 4, however, the multiplicity is not particularly limited in the system according to the invention.

Then, the text search is carried out of the whole document uniformly and so described hereinabove, however, information on the page boundary will be recorded in the text in a special symbol, and a page number successful in string matching can also be output as a matching result, the system of which is also included in the invention.

Further, the description has been given on an English text, however, the system can also be applied likewise to other languages.

Then, the text data is extracted through character recognition in the above example, however, the mode of a text content retrieval is apparently applicable to a text data inputted by hand, which is included in the invention.

Further, a system status has been described as illustrated in FIG. 4, however, it remains unchanged substantially in the case of miniature system or stand-alone system, which is also included in the invention. In particular, it is conceivable that a text file and an image file provided in another system be loaded to a small scale retrieval station, which is included in the invention.

Still further, it goes without saying that retrieval conditions can be combined through a logical operator or extended so as to retrieve the partial character string satisfying a relative positional relation. Particularly, a combina-

tive high retrieval can be realized at high speed through postprocessing by outputting a presence of each of a plurality of partial character strings.

As described above, according to the system of the invention, a desired document can be retrieved at high speed by referring to the contents of the document text, and also retrieved efficiently from a conception which is not conceivable at the point in time of having registered the document. Particularly at the time of registration, there is no necessity for worrying excessively about what is suitable to put as a classification code or key-word. A retrieval precision can be enhanced consequently and a noise occurrence can be suppressed at the same time.

Further, a text can be retrieved at high speed by juxtaposing the text search subsystem internally. A high speed operation can be attained particularly by adding a string matching circuit at every reading heads.

In the case of a retrieval for a large scale document file, the text contents can be retrieved by decreasing documents to be retrieved according to a keyword and bibliographical items, thus realizing an efficient retrieval as a whole.

Then, for obtaining a text data from document images, a document recognition result must have been inspected in each occasion to correct errors in the prior art, however, no attendant is particularly required therefor according to the invention. The text content retrieval has not been substantially realized hitherto for the reason mentioned above, but an effective text content retrieval can be secured by the invention.

What is claimed is:

1. A document storage and retrieval system for storing and retrieving textual documents, comprising:

image file means for storing textual documents which are digital image data, said textual documents including bibliographic items providing bibliographic information of said textual documents and body text data providing data of text found in bodies of said textual documents;

document recognition means, coupled to said image file means, for recognizing said textual documents, said document recognition means includes:

- (a) means for extracting pattern elements forming character patterns from said digital image data,
- (b) a document knowledge file for storing regulations of a layout of said bibliographic items in said textual documents as document knowledge,
- (c) character segmentation means for extracting character patterns by analyzing said pattern elements with reference to said document knowledge in said document knowledge file, and
- (d) recognition means for recognizing said extracted character patterns, said recognition means outputs a recognition result including said bibliographic items and said body text data with a layout structure name corresponding to the recognition result;

data base file means, coupled to said document recognition means, for storing said bibliographic items and information as bibliographic information of said outputted recognition result with said layout structure name;

text file means, coupled to said document recognition means, for storing at least said body text data as document contents of recognized textual documents;

input means for inputting a request of a search keyword;

retrieval means, coupled to said image file means, said data base file means, said text file means and said input

means, for retrieving digital image data of at least one textual document which includes said search keyword based on said stored bibliographic information and said stored body text data; and

output means, coupled to said retrieval means, for outputting said retrieved digital image data of at least one textual document.

2. A document storage and retrieval system according to claim 1, wherein said bibliographic items each include a title, an author's name or classification of a document.

3. A document storage and retrieval method for storing and retrieving textual documents, comprising the steps of: storing textual documents which are digital image data said textual documents including bibliographic items providing bibliographic information of said textual documents and body text data providing data of text found in bodies of said textual documents;

recognizing said textual documents, said recognizing step includes the steps of:

- (a) extracting pattern elements forming character patterns from said digital image data,
- (b) storing structural regulations of a layout of said bibliographic items in said textual documents as document knowledge,
- (c) extracting character patterns by analyzing said pattern elements with reference to said document knowledge, and
- (d) recognizing said extracted character patterns, and outputting a recognition result including said bibliographic items and said body text data with a layout structure name corresponding to the recognition result;

storing said bibliographic items and information as bibliographic information of said outputted recognition result with said layout structure name;

storing at least said body text data as document contents of recognized textual documents;

inputting a request of a search keyword;

retrieving digital image data of at least one textual document which includes said search keyword based on said stored bibliographic information and said stored body text data; and

outputting said retrieved digital image data of at least one document.

4. A document storage and retrieval method according to claim 3, wherein said bibliographic items each include a title, an author's name or classification of a document.

5. A document storage and retrieval system for storing and retrieving textual documents, comprising:

an image file storing textual document image data said textual documents including bibliographic items providing bibliographic information of said textual document image data and body text data providing data of text found in bodies of said textual documents image data;

means for extracting pattern elements forming character patterns from said textual document image data;

a document knowledge file storing structural regulations of a layout of bibliographic items in said textual document image data as document knowledge, according to each kind of textual document;

means for extracting subsets of pattern elements that constitute each bibliographic item, from said extracted pattern elements with reference to said document knowledge, and adding a name of a bibliographic item

corresponding to said extracted subset of pattern elements to said extracted subset of pattern elements;

means for recognizing character patterns as extracted pattern elements and generating a string of character codes corresponding to said extracted subset of pattern elements that constitutes a bibliographic item;

a text file storing said string of character codes when said string of character codes corresponds to document contents;

a data base file storing said string of character codes when said string of character codes corresponds to bibliographic information;

means for inputting a request of a search keyword; and

means for retrieving textual document image data of at least one textual document which includes a string of character codes corresponding to said search keyword based on strings of character codes stored in said text file and said data base file.

6. A document storage and retrieval system according to claim 5, wherein said bibliographic items are predetermined items of document attributes, including a title, an author's name and classification of a document.

7. A document storage and retrieval system according to claim 5, wherein said data base file stores strings of character codes corresponding to predetermined bibliographic items representing bibliographic information; and

wherein said text file stores strings of character codes corresponding to predetermined bibliographic items representing document contents.

8. A document storage and retrieval system according to claim 5, further comprising:

means for outputting a textual document image corresponding to said retrieved at least one textual document from said image file.

9. A document storage and retrieval system according to claim 5, further comprising:

a scanner reading an image of a textual document optically and generating said textual document image data.

10. In a document storage and retrieval system which holds data of textual documents in the form of an image and text, and retrieves textual document image data of at least one textual document which includes an inputted search keyword based on said data of documents in the form of text, a document storage method comprising the steps of:

reading textual document image data of textual documents in the form of an image, said textual document data including bibliographic items providing bibliographic information of said textual document image data and body text data providing data of text found in bodies of said textual documents image data;

extracting pattern elements forming character patterns from said textual document image data;

extracting subsets of pattern elements that constitute each of a plurality of said bibliographic items, from said extracted pattern elements, with reference to structural regulations of a layout of said bibliographic items in said textual document image data according to each kind of textual document;

adding a name of a bibliographic item corresponding to said extracted subset of pattern elements to said extracted subset of pattern elements;

recognizing character patterns as extracted pattern elements;

17

generating a string of character codes corresponding to
said extracted subset of pattern elements that constitute
a bibliographic item; and

storing strings of character codes in a text file when said 5
string of character codes corresponds to document
contents and in a data base file when said string of

18

character codes corresponds to bibliographic informa-
tion.

11. A document storage method according to claim 10,
comprising the step of:
reading an image of a document optically and generating
said document image data.

* * * * *